Extreme Learning Machine for Mammographic Risk Analysis

Wu, Wei; Shen, Qiang; Qu, Yanpeng; MacParthaláin, Neil Seosamh

Published in:
Proceedings of the 2010 UK Workshop on Computational Intelligence

Publication date:
2010

Citation for published version (APA):
Wu, W., Shen, Q., Qu, Y., & MacParthaláin, N. S. (2010). Extreme Learning Machine for Mammographic Risk Analysis. In *Proceedings of the 2010 UK Workshop on Computational Intelligence* IEEE Press.
http://hdl.handle.net/2160/5701

Extreme Learning Machine for Mammographic Risk Analysis

Yanpeng Qu, Qiang Shen, Neil Mac Parthaláin, and Wei Wu

***Abstract*— The assessment of mammographic risk analysis is an important issue in the medical field. Various approaches have been applied in order to achieve a higher accuracy in such analysis. In this paper, an approach known as Extreme Learning Machines (ELM), is employed to generate a single hidden layer neural network based classifier for estimating mammographic risk. ELM is able to avoid problems such as local minima, improper learning rate, and overfitting which iterative learning methods tend to suffer from. In addition the training phase of ELM is very fast. The performance of the ELM-trained neural network is compared with a number of state of the art classifiers. The results indicate that the use of ELM entails better classification accuracy for the problem of mammographic risk analysis.**

I. Introduction

Breast cancer is a major health issue, and perhaps the most common amongst women in the EU. It is estimated that between 8% and 13% of all women will develop breast cancer at some point during their lives [1], [2]. Furthermore, in the EU and US, breast cancer is acknowledged as the leading cause of death of women in their 40s [1], [2], [3]. Although increased levels of the occurrence of breast cancer have been recorded, so too has the level of early detection by screening using mammographic imaging and expert opinion. However, even expert radiologists sometimes fail to detect a significant proportion of mammographic abnormalities. In addition to this, a large number of detected abnormalities are usually discovered to be benign following medical investigation.

Existing mammographic Computer Aided Diagnosis (CAD) systems [4], [5] concentrate on the detection and classification of mammographic abnormalities. As breast tissue density increases however, the effectiveness of such systems in detecting such abnormalities is considerably reduced. Also, there is a strong correlation between mammographic breast tissue density and the risk of development of breast cancer. Automatic classification which has the ability to consider tissue density, and minimise human bias when searching for mammographic abnormalities is therefore highly desirable.

The extreme learning machine (ELM)-based classification approach was proposed in [6] with randomly assigned input weights and bias. Neural networks trained using ELM do not require adjustment of the input weights in the same way as with those using the backpropagation. Experimental results show that it works well when compared with backpropagation neural networks. A search of the ISI Web of Science has indicated that there are no current applications of ELM for mammographic analysis. This paper explores the potential of the ELM for mammographic risk analysis, an area where significant application studies have been reported [7], [8], [9].

Yanpeng Qu, Qiang Shen, and Neil Mac Parthaláin are with the Department of Computer Science, Llandinam Building, Aberystwyth University, Ceredigion, Wales, UK (email: {yyq09, qqs, ncm}@aber.ac.uk). Wei Wu is with School of Mathematical Sciences, Dalian University of Technology, Dalian, China (email: wuweiw@dlut.edu.cn).

The remainder of the paper is structured as follows. An outline of ELM for multi-class classification is presented in Section II. The data used for the experimental evaluation is described in Section III. The experimental results are discussed in Section IV. Section V concludes the paper, with further work pointed out.

II. ELM for Multi-class Classification

The ELM-based classifier implements multiple class recognition problems by employing a single-hidden layer feedforward neural network (SLFN) structure. For a dataset which contains N distinct objects: $(\mathbf{x}_i, \mathbf{t}_i)$, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T \in \mathbf{R}^n$ and $\mathbf{t}_i = [t_{i1}, t_{i2}, \ldots, t_{im}]^T \in \mathbf{R}^m$, the relationship between the (actual) outputs of the $SLFN$, with an infinite differentiable activation function $g(x)$, and the target outputs $\mathbf{t}_i$ is given by

$$\sum_{i=1}^{\widetilde{N}} \beta_i g(\mathbf{w}_i \cdot \mathbf{x}_j + b_i) = \mathbf{t}_j, \quad j = 1, \ldots, N. \tag{1}$$

Here, $\widetilde{N}$ is the number of hidden nodes, $\mathbf{w}_i = [w_{i1}, w_{i2}, \ldots, w_{in}]^T$ and $\beta_i = [\beta_{i1}, \beta_{i2}, \ldots, \beta_{im}]^T$ are the weight vectors connecting inputs to the ith hidden neuron and the ith hidden neuron to output neurons, respectively, and b_i is the bias of the ith hidden neuron.

Equation (1) can be written compactly as

$$\mathbf{H}\beta = \mathbf{T},$$

where

$$\mathbf{H} = \begin{bmatrix} g(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & g(\mathbf{w}_{\widetilde{N}} \cdot \mathbf{x}_1 + b_{\widetilde{N}}) \\ \vdots & \cdots & \vdots \\ g(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1) & \cdots & g(\mathbf{w}_{\widetilde{N}} \cdot \mathbf{x}_N + b_{\widetilde{N}}) \end{bmatrix}_{N \times \widetilde{N}}$$

is called the hidden layer output matrix of the neural network [6], and

$$\beta = \begin{bmatrix} \beta_1 & \cdots & \beta_{\widetilde{N}} \end{bmatrix}^T_{\widetilde{N} \times m}$$

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1 & \cdots & \mathbf{t}_N \end{bmatrix}^T_{N \times m}.$$

Traditionally, training of SLFN has typically applied the backpropagation learning algorithm to adjust the set of weights $(\mathbf{w}_i, \beta_i)$ and biases (b_i). It is common knowledge that the appropriate design of a backpropagation neural network is problem-dependent and, in most cases, a single hidden layer is sufficient (see [10] for example). However,

there is no clear guide to the determination of either the network structure or the parameters necessary to obtain an optimal neural network classifier. Although successful in many applications, several significant issues remain with backpropagation learning for neural networks:

1) The use of backpropagation requires the user to specify the value of the learning rate, whilst an improper setting for this rate will cause the algorithm to converge at a low speed or become unstable and divergent.
2) As a gradient descent-based method, learning with backpropagation may stop at a local minimum.
3) SLFN may be over-trained using backpropagation, resulting in poor generalisation and overfitting.
4) As with all gradient descent-based methods, learning using backpropagation can be time consuming.

In addressing these issues, it has been established that for SLFNs with additive or RBF hidden nodes, the hidden node parameters may be randomly specified initially. The output weights can then be analytically determined when the are used to approximate any continuous target function [6]. Also, it is shown that the upper bound of the required number of hidden nodes is the number of distinct training objects (i.e. $\widetilde{N} \leq N$). Thus, given (a pre-specified) $\widetilde{N}$, associated with parameters $(\mathbf{w}_i, b_i)$, the hidden nodes can be randomly generated. Determining the output weights β is as simple as finding the least-square solutions to the given linear system.

$$\widehat{\beta} = \mathbf{H}^{\dagger}\mathbf{T},$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalised inverse [11], [12] of the hidden layer output matrix $\mathbf{H}$. Such $\widehat{\beta}$ has following important properties [6]:

1) *Minimum training error.* The special solution $\widehat{\beta} = \mathbf{H}^{\dagger}\mathbf{T}$ is one of the least-squares solutions of a general linear system $\mathbf{H}\beta = \mathbf{T}$, meaning that the smallest training error can be reached by this special solution:

$$\|\mathbf{H}\widehat{\beta} - \mathbf{T}\| = \|\mathbf{H}\mathbf{H}^{\dagger}\mathbf{T} - \mathbf{T}\| = \min_{\beta} \|\mathbf{H}\beta - \mathbf{T}\|.$$

Although almost all learning algorithms are intended to reach the minimum training error, however, most of them can not reach it because of local minimum or infinite training iterations is usually not allowed in applications.

2) *Smallest norm of weights.* Further, the special solution $\widehat{\beta} = \mathbf{H}^{\dagger}\mathbf{T}$ has the smallest norm among all least-squares solutions of $\mathbf{H}\beta = \mathbf{T}$:

$$\|\widehat{\beta}\| = \|\mathbf{H}^{\dagger}\mathbf{T}\| \leq \|\beta\|,$$
$$\forall \beta \in \left\{\beta : \|\mathbf{H}\beta - \mathbf{T}\| \leq \|\mathbf{H}\mathbf{z} - \mathbf{T}\|, \forall \mathbf{z} \in \mathbf{R}^{\widetilde{N}\times N}\right\}.$$

3) The minimum norm least-squares solution of $\mathbf{H}\beta = \mathbf{T}$ is unique, which is $\widehat{\beta} = \mathbf{H}^{\dagger}\mathbf{T}$.

Calculation of the weight between hidden layer and output layer is done in a single step. This avoids any lengthy training procedure where the network parameters are adjusted iteratively.

Following the above discussion, a three-step ELM algorithm can be summarised as follows:

ELM($\mathbb{N}$,g,$\widetilde{N}$)
$\mathbb{N}$, the training set $\{(\mathbf{x}_i, \mathbf{t}_i)|\mathbf{x}_i \in \mathbf{R}^n, \mathbf{t}_i \in \mathbf{R}^m, i = 1, \cdots, N\}$,
g, the activation function,
$\widetilde{N}$, the number of hidden nodes.

(1) Randomly assign hidden node parameters $(\mathbf{w}_j, b_j)$, $i = 1, \cdots, \widetilde{N}$,
(2) Calculate the hidden layer output matrix $\mathbf{H}$,
(3) Calculate the output weight β.

Fig. 1. The ELM Algorithm

III. Experimental Data

The data used for the experimental evaluation in this paper is derived from features extracted from images in the Mammographic Image Analysis Society (MIAS) database [13]. Medio-Lateral-Oblique (MLO) left and right mammograms of 161 women (322 objects). Each data object or mammogram is represented by 280 features, 10 which relate to morphological characteristics, and the remaining 270 from the extracted image texture information. The spatial resolution of the images is $50\mu m \times 50\mu m$ and quantised to 8 bits with a linear optical density in the range 0–3.2.

The class labels for each mammogram are assigned using the consensus opinion (via majority voting) of three expert radiologists. Figure 2 shows 4 mammograms covering a range of breast tissue density [14]. Each of these images represents a different BI-RADS class. The American College of Radiology BI-RADS [15] is a widely used risk assessment model. It aims to classify a mammogram into one of four classes according to breast density. The classes can be explained as follows. BI-RADS I: an almost entirely fatty breast, not dense; BI-RADS II: some fibroglandular tissue is present; BI-RADS III: the breast is heterogeneously dense; BI-RADS IV: the breast is extremely dense. Although BI-RADS is becoming a radiological standard, other risk assessment models exist that aim to classify breasts according to different aspects or features present in the mammogram [16].

IV. Experimental Results

A. Experimental Set-up

The set-up employed for the experimental evaluation in this paper is shown in Figure 3. Note that the feature extraction technique employed here is that which is used in [14]. The initial stages of this feature extraction technique involve the segmentation and filtering of the mammographic images: all mammograms are pre-processed to identify the breast region and remove image background, labels, and pectoral muscle areas. The segmentation step results in a very minor loss of skin-line pixels in the breast area, however these

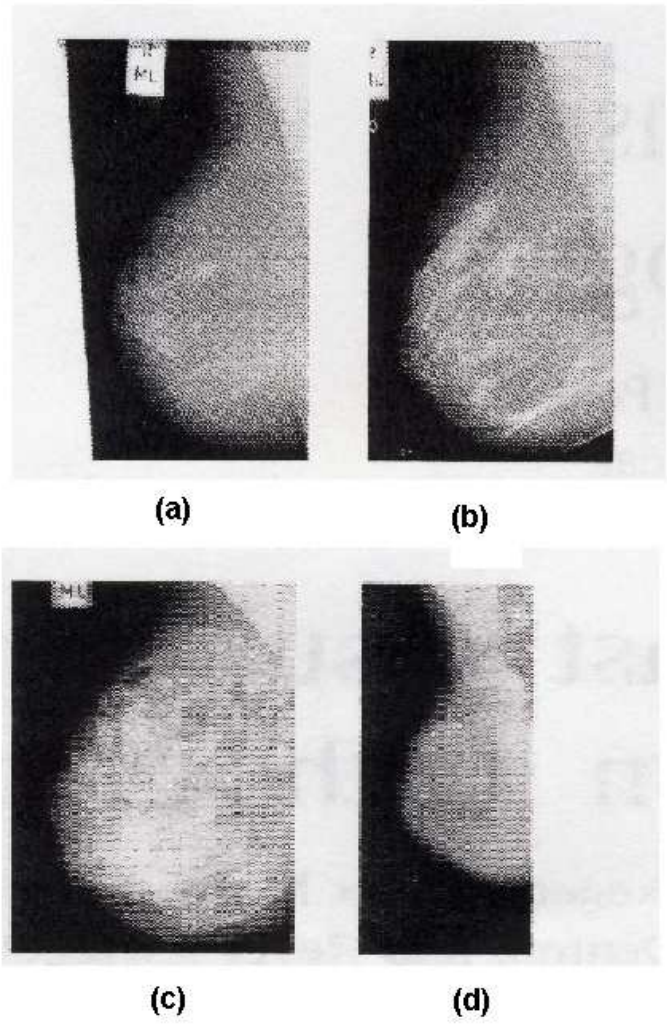

Fig. 2. Mammograms showing 4 different breast densities ranging from low density (a) to high density (d).

pixels are not required for tissue density estimation. Then, a feature extraction step is performed, where the fuzzy c-means (FCM) algorithm is employed which results in the division of the breast into two clusters. A co-occurrence matrix (which is essentially a 2D histogram) is then used to derive a feature set which results in 10 features to describe morphological characteristics and 216 for the texture information (226 total). This feature set is then labelled using the consensus opinion of 3 experts to assign a label to each object mammogram using the BI-RADS [15] classification.

The design of the ELM-based learning requires the setting of one user-defined parameter: the number of nodes in the hidden layer. A series of experiments were carried out in order to ascertain the variation in the resulting classification accuracy by changing $\tilde{N}$, with it ranging from 10 to 322. Figure 4 shows the relationship between $\tilde{N}$ and the resulting classification accuracy. Since ELM-based learning is stochastic, the variance over 10 runs has also been included and is represented by the error bars. Note that the accuracy begins to decline when the number of nodes is greater than 50. The computational cost also increases if more hidden nodes are used. Thus, in Figure 4, only the results obtained using 50 hidden nodes are analysed. Note that leave-one-out cross validation (LOOCV) is employed for model selection such that the results can be compared with existing work.

B. Performance Evaluation

To evaluate the performance of applying ELM to train SLFN on the mammographic dataset, the experimental results are compared with those obtained by fuzzy-rough nearest-neighbour FRNN [17], fuzzy nearest-neighbour (FNN) [18], vaguely quantified nearest-neighbour (VQNN) [19], and other popular classifiers. Classification accuracy, standard deviation (std dev) and confusion matrices are used to support such comparison, as shown in Table I. As can be seen, a classification accuracy of 73.91% and std dev of 2.0135 were achieved by ELM in comparison to 69.90% and 45.84 by FRNN, 71.75% and 45.43 by VQNN, 62.42% and 48.39 by FNN, 66.78% and 47.73 by JRip and 63.98% and 47.97 by PART. The results suggest that the ELM-trained network gives the best and most stable performance. The results for each of the classifiers were also compared statistically using a paired t-test. The result returned by the ELM trained classifier was not statistically insignificant, while performing considerably better than some learners, e.g. FNN.

Perhaps most important however is that ELM-based classifier manages to reduce the class confusion between classes II and III. Indeed, if these results are compared with those of [14], it can be seen that the work here offers a considerable improvement in the ability to distinguish between classes II and III. It should be noted that the work of [14] is the current state-of-the-art for mammographic risk assessment.

V. Conclusion

This paper has presented an effective classification method for Mammographic risk analysis. Its performance has been compared with state-of-the-art classifier learning methods. The ELM-based neural network approach can perform the multi-category classification directly, without any modification of the initial parameter settings. Experimental results show that the ELM-trained classifier achieves a higher classification accuracy than other algorithms. Also, unlike a backpropagation neural network, the performance of ELM is affected by only one user-defined parameter, which can be determined through trial-and-error for a particular dataset with an identified upper bound.

Note that in this paper, work is focused on classification. The effect of feature selection on the learned classifiers is not investigated. A further extension to this research would be to explore how ELM-trained classifiers would perform following dimensionality reduction using various feature selection methods. This could form the basis for an integrated learning framework approach which takes advantage of the improved performance of ELM.

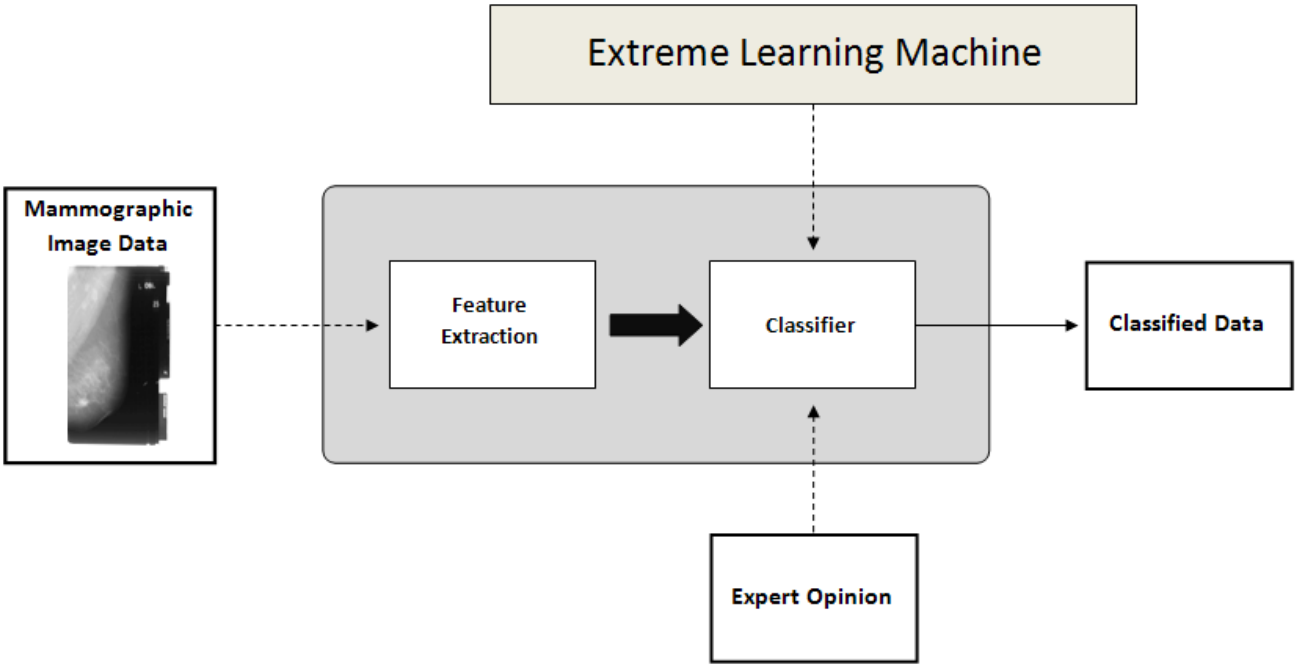

Fig. 3. Experimental set-up for ELM classification

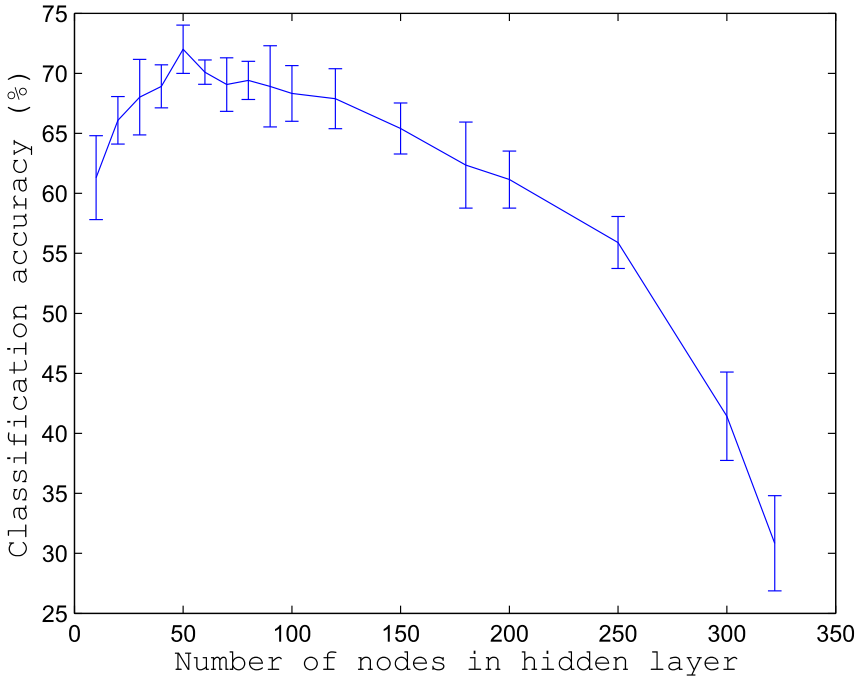

Fig. 4. Number of nodes in hidden layer vs. classification accuracy

TABLE I

CONFUSION MATRICES, CLASSIFICATION ACCURACIES AND STANDARD DEVIATION FOR THE MIAS DATASET CLASSIFICATION USING ELM AND OTHER FIVE DIFFERENT CLASSIFIER LEARNERS

ELM
(*Classification accy* $= 73.91\%$, *std dev* $= 2.0135$)

	I	II	III	IV
I	74	10	2	1
II	12	73	17	1
III	2	18	72	3
IV	1	3	14	19

FRNN
(*Classification accy* $= 69.90\%$, *std dev* $= 45.84$)

	I	II	III	IV
I	74	12	1	0
II	12	67	22	2
III	0	26	64	5
IV	2	3	12	20

VQNN
(*Classification accy* $= 71.75\%$, *std dev* $= 45.43$)

	I	II	III	IV
I	74	11	1	1
II	13	68	20	2
III	0	22	70	3
IV	2	2	14	19

FNN
(*Classification accy* $= 62.42\%$, *std dev* $= 48.39$)

	I	II	III	IV
I	58	20	9	0
II	16	67	20	0
III	2	29	60	4
IV	1	2	18	16

JRip
(*Classification accy* $= 66.78\%$, *std dev* $= 47.73$)

	I	II	III	IV
I	72	10	4	1
II	15	65	21	2
III	2	30	56	7
IV	0	4	11	22

PART
(*Classification accy* $= 63.98\%$, *std dev* $= 47.97$)

	I	II	III	IV
I	64	17	4	2
II	19	64	17	3
III	5	25	56	9
IV	3	2	10	22

Another important area for future work includes the in-depth investigation of the level of agreement between individual expert classification of mammographic risk for each image and those obtained by the ELM classifier.

Finally, a more complete comparison of ELM with other techniques for classifier learning, over different datasets from other application domains, would form the basis for a wider series of topics for future studies.

REFERENCES

[1] F. Bray, P. McCarron, and D. Parkin, "The changing global patterns of female breast cancer incidence and mortality," *Breast Cancer Research*, vol. 6, pp. 229–239, 2004.

[2] Eurostat, "Health statistics atlas on mortality in the european union," *Official Journal of the European Union*, 2002.

[3] S. Buseman, J. Mouchawar, N. Calonge, and T. Byers, "Mammography screening matters for young women with breast carcinoma," *Cancer*, vol. 97(2), pp. 352–358, 2003.

[4] iCAD Second Look, http://www.icadmed.com. *accessed*: 10/04/2010.

[5] R2 Image Checker, http://www.r2tech.com. *accessed*: 10/04/2010.

[6] G. Huang, Q. Zhu, and C. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70(1-3), pp. 489–501, 2006.

[7] M. Brady and R. Highnam, *Mammographic Image Analysis*. Kluwer Series on Medical Image Understanding, 1999.

[8] A. Hassanien, "Fuzzy rough sets hybrid scheme for breast cancer detection," *Image and Vision Computing*, vol. 25(2), pp. 172–183, 2007.

[9] M. Roffilli, "Advanced machine learning techniques for digital mammography." Technical Report UBLCS-2006-12. University of Bologna (Italy), 2006.

[10] J. A. Baker, P. J. Kornguth, J. Lo, M. E. Williford, and C. E. Floyd, "Breast cancer: prediction with artificial neural network based on BIRADS standardized lexicon," *Radiology*, vol. 196(3), pp. 817–822, 1995.

[11] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and its Applications*. New York: Wiley, 1971.

[12] D. Serre, *Matrices: Theory and Applications*. Springer, 2002.

[13] J. Suckling, J. Partner, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, D. B. P. Taylor, and J. Savage, "The mammographic image analysis society digital mammogram database," in *International Workshop on Digital Mammography*, pp. 211–221, 1994.

[14] A. Oliver, J. Freixenet, R. Marti, J. Pont, E. Perez, E. Denton, and R. Zwiggelaar, "A novel breast tissue density classification methodology," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12(1), pp. 55–65, 2008.

[15] American College of Radiology, *Illustrated Breast Imaging Reporting and Data System BIRADS*, 3rd ed., 1998.

[16] H. Strange and R. Zwiggelaar, "Classification performance related to intrinsic dimensionality in mammographic image analysis," in *Proceedings of the Thirteenth Annual Conference on Medical Image Understanding and Analysis*, pp. 219–223, 2009.

[17] N. MacParthaláin, R. Jensen, Q. Shen, and R. Zwiggelaar, "Fuzzy-rough approaches for mammographic risk analysis," *Intelligent Data Analysis*, vol. 14(2), pp. 225–244, 2010.

[18] R. Jensen and C. Cornelis, "A new approach to fuzzy-rough nearest neighbour classification," in *Proceedings of the 6th International Conference on Rough Sets and Current Trends in Computing*, pp. 310–319, 2008.

[19] C. Cornelis, M. D. Cock, and A. Radzikowska, "Vaguely quantified rough sets," *Lecture Notes in Artificial Intelligence*, vol. 4482, pp. 87–94, 2007.